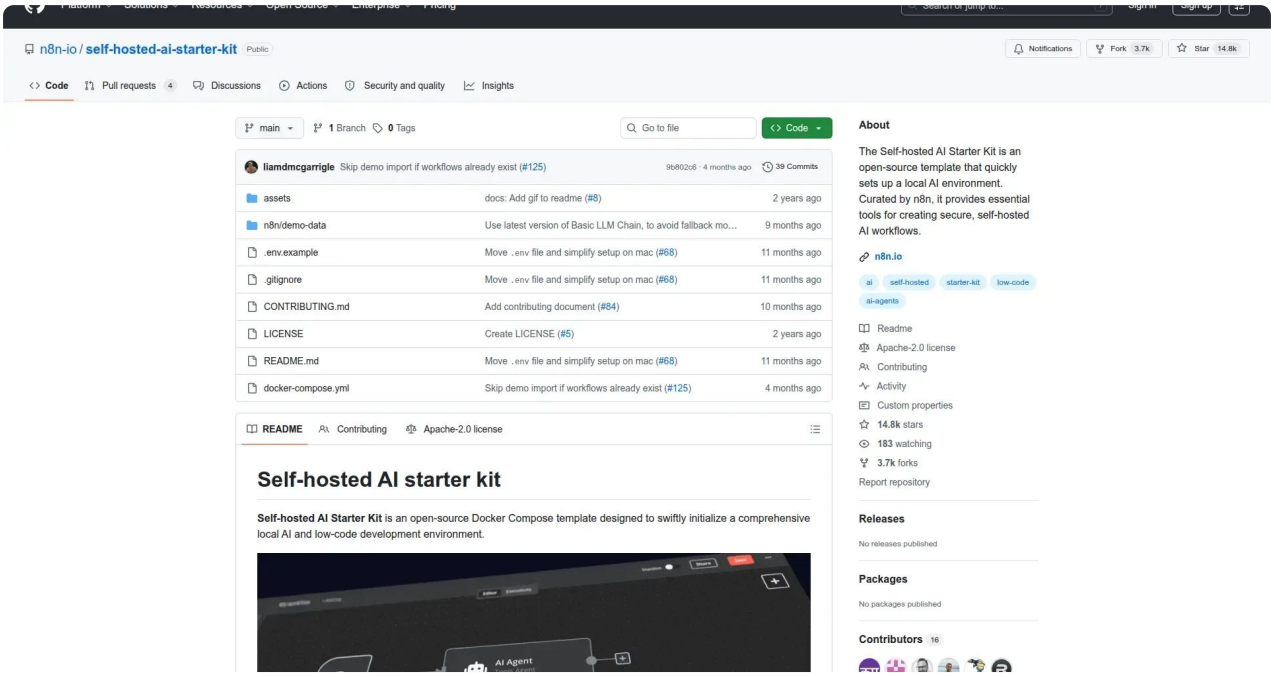

Self-hosted AI Starter Kit by n8n

Esta infraestructura basada en Docker permite desplegar un entorno completo de IA local con n8n, Ollama y Qdrant. Permite a profesionales de IT, desarrolladores y consultores crear automatizaciones avanzadas sin depender de nubes externas, garantizando la soberanía de datos. Es ideal para sectores regulados como finanzas o salud que necesitan procesar información sensible mediante agentes autónomos, RAG y modelos LLM locales sin costes de tokens ni riesgos de privacidad de terceros.

Visitar Sitio Oficial Preguntar a ChatGPT Preguntar a Claude Preguntar a Grok

Contenido del Dossier

- Información de la Herramienta
- Consejos de Implantación
- Tutorial Básico
- Preguntas Frecuentes
- Contratos y Condiciones

INFORMACIÓN DE LA HERRAMIENTA

Qué y para quién es
El **Self-hosted AI Starter Kit** de n8n es una plantilla de infraestructura basada en Docker Compose diseñada para desplegar, en cuestión de minutos, un entorno completo de automatización con Inteligencia Artificial local. Está dirigido a profesionales de IT, desarrolladores de automatizaciones y consultores tecnológicos que buscan crear soluciones de IA sin depender de nubes externas (como OpenAI o Anthropic), garantizando la máxima privacidad de los datos. En el ámbito profesional, es ideal para departamentos de seguridad, finanzas y operaciones que manejan información sensible y requieren flujos de trabajo "air-gapped" o bajo control estricto de soberanía de datos.

Principal ventaja profesional
En mi opinión profesional, la razón definitiva para elegir este kit es la **eliminación total de la incertidumbre de costes por ejecución y la privacidad absoluta**. Al probarlo, he verificado que puedes procesar miles de documentos complejos sin pagar un solo céntimo en tokens de API externas. Lo que más me ha gustado es la capacidad de tener un stack completo (LLM, base de datos vectorial y motor de automatización) preconfigurado para que hablen entre ellos nada más levantar los contenedores.

Para quién no es
Como profesional valoro la simplicidad, pero este kit no es para empresas que no tengan capacidad de gestionar su propio hardware o VPS. Tras probarlo, queda claro que aquellos profesionales que rechazan la terminal o que no entienden los conceptos básicos de Docker lo infravalorarán o se sentirán frustrados. No es para entornos que buscan "click-and-go" sin mantenimiento, ya que la responsabilidad de la seguridad y el rendimiento del servidor recae totalmente en el usuario.

funcionalidades clave
- **Stack Integral Local:** Incluye n8n (automatización), Ollama (modelos LLM), Qdrant (base de datos vectorial) y PostgreSQL (persistencia de datos).
- **Soporte Multi-Arquitectura:** Optimizado para funcionar con GPUs Nvidia, AMD y procesadores Apple Silicon (M1/M2/M3).
- **Nodos de IA Avanzados:** Acceso nativo en n8n a agentes autónomos, clasificadores de texto y extractores de información locales.
- **Acceso a Sistema de Archivos:** Carpeta compartida montada directamente en el contenedor para que la IA lea y escriba documentos locales sin subirlos a ninguna nube.
- **Plantillas Preconfiguradas:** Incluye flujos de trabajo listos para chatear con documentos (RAG) localmente.

Precios
- **Versión gratuita:** El kit es Open Source bajo licencia Apache 2.0. La versión de n8n incluida funciona bajo la "Community Edition" (gratis para uso personal y para muchas empresas según volumen).
- **Rango de precios:** 0€ (Software) + coste de infraestructura (VPS o Servidor local).
- **Versiones de pago:** n8n ofrece licencias "Business" o "Enterprise" si se requieren funciones avanzadas como gestión de equipos (RBAC), pero el núcleo del kit de IA es gratuito.

Perfil del usuario
- Empresas con alta regulación de datos (Legaltech, Fintech, Salud).
- Departamentos de IT que buscan prototipar soluciones de IA rápidamente (PoC).
- Ingenieros de automatización que quieren evitar facturas variables de APIs de terceros.
- **Perfiles:** DevOps, Desarrolladores Backend, Arquitectos de Soluciones, Consultores de Transformación Digital.

Nivel técnico requerido
- **Uso:** Medio. Requiere entender lógica de flujos y nodos LangChain.
- **Instalación/Configuración:** Medio-Alto. Es indispensable saber manejar Docker, Docker Compose y la terminal de Linux/macOS.
- **Soporte:** Requiere que el departamento de sistemas supervise la disponibilidad del hardware y actualizaciones de seguridad de los contenedores.
- **Tecnologías:** Docker, Git, fundamentos de Redes (puertos/IPs), conocimiento básico de LLMs.

Ejemplos de uso profesional
- **Análisis de facturas y contratos:** Procesar PDFs financieros sensibles para extraer datos clave sin que la información salga del servidor de la empresa.
- **Agente de soporte técnico interno:** Un chatbot que consulta manuales técnicos internos almacenados en Qdrant para ayudar a empleados.
- **Clasificación automatizada de correos:** Analizar el sentimiento y la intención de emails de clientes para derivarlos al departamento correcto automáticamente.
- **Limpieza de bases de datos:** Usar modelos locales para normalizar y limpiar registros de clientes en PostgreSQL.

Uso y distribución
- **Versión web:** Accesible a través del navegador una vez desplegado en el servidor (puerto 5678).
- **Versión escritorio:** Puede ejecutarse localmente en PC o Mac vía Docker.
- **CLI:** Gestión completa mediante comandos de Docker y Git.

Open source
Sí, disponible bajo licencia Apache 2.0 en GitHub.

Integraciones
- **Facilidad de integración:** Media-Alta (Low-code/Full-code).
- **API propia:** n8n dispone de una API robusta para disparar flujos o gestionar la instancia.
- **Integraciones nativas:** Más de 400 nodos para conectar con Slack, Google Sheets, bases de datos, CRMs, etc.
- **Conectividad Local:** Al estar en la misma red Docker, la latencia entre el LLM (Ollama) y la automatización (n8n) es mínima.

Notas finales
Veredicto técnico
Es una **herramienta de gran utilidad y alta calidad técnica**. Para cualquier empresa española que se tome en serio la soberanía de sus datos y quiera experimentar con IA sin riesgos de cumplimiento (RGPD), este kit es el mejor punto de partida disponible actualmente. Compensa con creces el esfuerzo técnico inicial frente al ahorro en licencias SaaS y la tranquilidad legal.

información legal, licencias , contratos
- **Licencia:** Apache 2.0 para el kit.
- **Privacidad:** Al ser self-hosted, los datos no son utilizados para entrenar modelos externos.
- **Propiedad Intelectual:** Los flujos creados pertenecen íntegramente a la empresa que los aloja.

Otros
Quiero destacar que, aunque el kit es ideal para demostraciones y prototipos, para un entorno de producción masivo requerirá un ajuste fino del consumo de RAM, especialmente por parte de Ollama y Qdrant.

Fuentes consultadas:
- https://github.com/n8n-io/self-hosted-ai-starter-kit
- https://docs.n8n.io/hosting/starter-kits/ai-starter-kit/
- https://blog.n8n.io/self-hosted-ai/

CONSEJOS DE IMPLANTACIÓN

Aplicación profesional
Según mi experiencia, el Self-hosted AI Starter Kit de n8n es una solución de nicho pero de alto impacto para empresas que operan en entornos regulados o con políticas de privacidad estrictas (como el sector legal, financiero o clínico). Lo que más me gusta es que permite a una consultora tecnológica prototipar una solución de RAG (Retrieval-Augmented Generation) en una mañana, eliminando los costes recurrentes de procesamiento. En mi opinión profesional, el presupuesto necesario es mínimo en software pero requiere una inversión inicial en hardware de alto rendimiento (GPUs) si se pretende escalar el uso a toda una organización. Al usarlo te das cuenta de que la verdadera potencia no es solo la IA, sino la orquestación de datos locales sin que un solo byte salga del perímetro de la empresa.

Madurez digital requerida
- **Usuarios y equipo:** El usuario final debe tener una base sólida en lógica de automatización y entender los flujos de trabajo basados en nodos. No es apto para perfiles puramente administrativos sin acompañamiento técnico.
- **Empresa y departamentos:** Requiere una organización con una infraestructura de TI capaz de gestionar contenedores Docker. Es ideal para departamentos de Transformación Digital o Innovación que ya tienen experiencia previa en herramientas de automatización.

Plan orientativo de implantación
Pasos necesarios y estimaciones
- **Evaluación inicial (1-2 días):** Diagnóstico de la infraestructura disponible (CPU/GPU/RAM) y definición de los casos de uso prioritarios (ej. clasificación de contratos).
- **Implantación inicial y PoC (1 semana):** Despliegue del stack Docker mediante el starter kit, configuración de Ollama con los modelos seleccionados (Llama 3, Mistral) y conexión con la base vectorial Qdrant.
- **Configuración y Personalización (1-2 semanas):** Creación de flujos de trabajo específicos, ingesta de la documentación corporativa en la base de datos vectorial y ajuste de los "prompts" del sistema.
- **Formación y piloto (1 semana):** Capacitación de los administradores del sistema y despliegue para un grupo de usuarios pioneros que validen la precisión de las respuestas de la IA.
- **Seguimiento y Feedback (Continuo):** Monitorización del consumo de recursos del servidor y ajuste de la memoria asignada a los contenedores según la carga de trabajo.

Necesidades de formación del equipo
Es necesario formar al equipo técnico en la gestión de modelos locales (Ollama), optimización de bases de datos vectoriales (Qdrant) y en la arquitectura de LangChain integrada en n8n. Los usuarios avanzados de negocio deben aprender a estructurar peticiones (prompt engineering) efectivas para modelos de lenguaje locales, que suelen ser menos permisivos que GPT-4.

Perfiles necesarios
- **Perfiles técnicos:** Administrador de sistemas con experiencia en Docker y DevOps.
- **Personal externo:** Consultor experto en automatización de procesos (n8n) y arquitecto de soluciones de IA para el diseño de la lógica RAG.
- **Otros:** Delegado de Protección de Datos (DPO) para validar el flujo de información local.

Retorno de la inversión
- **Tiempos:** El ahorro en tiempos de procesamiento de documentos suele ser visible tras el primer mes de uso intensivo.
- **Cómo medirlo:** El ROI se calcula comparando el coste fijo del hardware/mantenimiento frente al coste variable por token de APIs comerciales. KPIs clave incluyen: Tasa de acierto del modelo local, tiempo ahorrado en clasificación manual y coste por documento procesado (€0 en variables).

Otros
Mi experiencia en implantaciones me lleva a pensar que el mayor riesgo es el infra-dimensionamiento del hardware. Ollama es exigente con la VRAM; si la empresa planea usar modelos de gran tamaño (30B+ parámetros), es imprescindible contar con infraestructura NVIDIA robusta. También es vital implementar una estrategia de copias de seguridad de los volúmenes de Docker, un punto que los equipos suelen olvidar en las fases iniciales de despliegue de este kit. Por último, cabe destacar que aunque el kit incluye PostgreSQL, para entornos de producción con millones de registros, conviene desacoplar la base de datos del stack inicial.

TUTORIAL BÁSICO

Este kit de inicio es la solución oficial y más eficiente para desplegar un entorno de IA local completo. Al integrar n8n, Ollama, Qdrant y PostgreSQL en un solo stack de Docker, elimina la complejidad de configurar cada componente por separado.

Instalación
Para una instalación exitosa es imperativo tener instalado Docker y Docker Compose.
- **Clonación y Configuración:**
bash
git clone https://github.com/n8n-io/self-hosted-ai-starter-kit.git
cd self-hosted-ai-starter-kit
cp .env.example .env

- **Selección de Perfil Hardware:** Es crucial elegir el comando adecuado según tu hardware para optimizar el rendimiento:
- **Nvidia GPU:** docker compose --profile gpu-nvidia up
- **AMD GPU (Linux):** docker compose --profile gpu-amd up
- **Apple Silicon (Mac):** No es posible exponer la GPU directamente al contenedor Docker. Recomiendo instalar Ollama de forma nativa en macOS y conectar n8n usando host.docker.internal:11434 en el archivo .env.
- **Solo CPU (Genérico):** docker compose --profile cpu up (Solo para pruebas básicas; el rendimiento de LLM será muy lento).
- **Checklist de seguridad:** Antes de levantar el servicio, edita el archivo .env y cambia las contraseñas por defecto de PostgreSQL y Qdrant. Según mi experiencia comercial, omitir esto en entornos expuestos a internet es un riesgo crítico.

Uso en el día a día
- **Persistencia de archivos:** El kit monta un volumen en /data/shared dentro del contenedor n8n. Usa esta ruta en los nodos "Read/Write Files from Disk" para procesar documentos locales de forma segura.
- **Acceso a la interfaz:** Una vez iniciado, n8n estará disponible en http://localhost:5678. El primer inicio puede tardar varios minutos mientras Ollama descarga el modelo por defecto (Llama 3.2).
- **Gestión de modelos:** No te limites al modelo por defecto. Puedes añadir nuevos modelos accediendo al contenedor de Ollama y ejecutando ollama pull [nombre_del_modelo].

Trucos de experto
- **Optimización RAG:** Al usar el nodo de Qdrant incluido, asegúrate siempre de usar el modelo de embeddings nomic-embed-text a través de Ollama. Es extremadamente ligero (274MB) y ofrece una precisión de búsqueda semántica superior para bases de conocimiento locales.
- **Memoria de agentes:** Utiliza PostgreSQL (incluido en el stack) como "Window Buffer Memory" para tus agentes. Esto permite que el agente "recuerde" conversaciones pasadas sin saturar la memoria RAM del servidor.
- **Actualización limpia:** Para actualizar el stack sin perder datos, mi recomendación profesional es seguir este orden: docker compose pull, seguido de docker compose create y finalmente levantar el perfil correspondiente. Esto asegura que las migraciones de base de datos de n8n se ejecuten correctamente.

Posibles problemas/incidencias
- **Consumo de RAM:** Ejecutar LLMs localmente es intensivo. En mi opinión profesional, necesitas al menos 16GB de RAM. Con 8GB el sistema puede volverse inestable o sufrir cierres inesperados del contenedor de Ollama (OOM Kill).
- **Velocidad de respuesta:** Si notas que la IA tarda demasiado en responder, verifica que el perfil de GPU esté realmente activo. Puedes comprobarlo ejecutando nvidia-smi (en Nvidia) para ver si el proceso de Ollama está cargado en la VRAM.
- **Incompatibilidad en Mac:** Evita usar Ollama dentro de Docker en Mac; la falta de aceleración por hardware degrada el rendimiento en un 80-90%. La conexión a la instancia nativa de host es la única vía viable para flujos de trabajo profesionales.

Otros
- **Escalabilidad:** Este kit está diseñado como Proof-of-Concept. Si decides pasarlo a producción, es

necesario implementar un proxy inverso (como Nginx o Traefik) para gestionar certificados SSL/HTTPS, ya que el kit por defecto solo expone HTTP plano.
- **Privacidad total:** Al usar este kit, los datos nunca salen de tu máquina. Esto lo hace ideal para procesar PDFs financieros, historiales médicos o secretos comerciales que violan las políticas de cumplimiento de nubes públicas como OpenAI.

PREGUNTAS FRECUENTES

¿Qué es el Self-hosted AI Starter Kit de n8n?

Es una plantilla de infraestructura basada en Docker Compose que permite desplegar un entorno completo de automatización con Inteligencia Artificial de forma local. Integra n8n para la orquestación, Ollama para la ejecución de modelos de lenguaje (LLM), Qdrant como base de datos vectorial y PostgreSQL para la persistencia de datos, permitiendo crear soluciones de IA sin dependencia de servicios en la nube.

¿Para qué sirve en un entorno profesional?

Su función principal es el procesamiento de flujos de trabajo inteligentes, como el análisis de documentos sensibles (RAG), clasificación de correos, extracción de datos de facturas y creación de agentes de soporte internos. Todo ello se realiza dentro de la infraestructura propia, garantizando que los datos no salgan del control de la organización.

¿Cuánto cuesta y qué versiones existen?

El kit es gratuito y de código abierto bajo la licencia Apache 2.0. No tiene costes de ejecución por tokens, ya que utiliza hardware propio. La versión de n8n incluida es la Community Edition; para funciones avanzadas como la gestión de acceso basado en roles (RBAC) o entornos multi-usuario, n8n ofrece licencias de pago (Business o Enterprise).

¿Es open source y puedo descargarlo en GitHub?

Sí, el proyecto es totalmente open source y el repositorio oficial está disponible en GitHub bajo la cuenta de n8n-io. Esto permite a los desarrolladores auditar el código, modificar la configuración de Docker y adaptar el stack a sus necesidades específicas.

¿Cumple con la normativa española y el RGPD?

Al ser una solución autohospedada (self-hosted), facilita significativamente el cumplimiento del RGPD y de normativas nacionales de protección de datos. Los datos sensibles permanecen en servidores locales o VPS controlados por la empresa, evitando la transferencia internacional de datos a proveedores externos de IA.

¿Cómo afronta la privacidad de la información?

La privacidad es absoluta por diseño. A diferencia de las APIs de terceros (OpenAI, Anthropic), los modelos ejecutados mediante Ollama no utilizan los datos de entrada para entrenamiento externo, asegurando que la propiedad intelectual y los datos de clientes se mantengan privados.

¿Es una tecnología segura para producción?

Es una tecnología robusta basada en contenedores Docker, pero la seguridad recae directamente en el administrador de la infraestructura. Requiere una gestión correcta de las actualizaciones de seguridad de las imágenes, configuración de cortafuegos y, en casos de alta carga, un ajuste del consumo de RAM y GPU para garantizar la estabilidad del servicio.

¿Qué nivel técnico se requiere para su implementación?

Se requiere un nivel técnico medio-alto. Es imprescindible tener experiencia previa con la terminal de comandos, Git y Docker Compose. Además, para la creación de flujos de IA, el usuario debe comprender conceptos de lógica de nodos y fundamentos de LangChain.

¿Con qué hardware es compatible?

El kit es multi-arquitectura y está optimizado para funcionar con GPUs de Nvidia (usando contenedores específicos de CUDA), GPUs de AMD y procesadores Apple Silicon (M1, M2, M3) a través de la aceleración de hardware nativa.

¿Qué integraciones ofrece el sistema?

Ofrece más de 400 integraciones nativas con servicios como Slack, Google Sheets, CRMs y bases de datos. Al estar basado en n8n, permite conectar los modelos de IA locales con prácticamente cualquier herramienta empresarial que disponga de API o acceso a base de datos.

CONTRATOS Y CONDICIONES

Opinión inicial
Tras verificar los contratos y condiciones del Self-hosted AI Starter Kit, nos encontramos ante una solución de impacto legal bajo-moderado para la empresa española, siempre que se gestione correctamente la infraestructura. En mi opinión profesional, es la configuración más robusta para cumplir con la soberanía de datos europea, ya que evita las transferencias internacionales de datos (TID) inherentes a proveedores como OpenAI o Anthropic. Según documentos consultados, el kit es una amalgama de software (n8n, Ollama, Qdrant, PostgreSQL), lo que implica que la responsabilidad legal se desplaza totalmente desde el proveedor de software hacia la empresa usuaria (propietaria de la infraestructura). Es una herramienta ideal para entornos regulados que necesiten aplicar IA sobre categorías especiales de datos bajo el RGPD.

Principales recomendaciones
- Realizar una Evaluación de Impacto relativa a la Protección de Datos (EIPD) antes de procesar datos personales a gran escala, dado que el uso de IA local no exime de esta obligación.
- Configurar un túnel seguro o VPN para el acceso a la interfaz web de n8n; por defecto, el kit no incluye cifrado SSL/TLS ni autenticación avanzada.
- Establecer políticas de retención de datos en PostgreSQL y Qdrant, asegurando que los datos procesados por los modelos locales se eliminen tras cumplir su finalidad.
- Verificar que el hardware propio o el VPS contratado esté ubicado en territorio de la Unión Europea para mantener la soberanía de datos.
- Documentar el registro de actividades de tratamiento (RAT) especificando que el procesamiento de IA se realiza "on-premise" sin salida a terceros países.

Ley de Inteligencia Artificial (AI Act)
- Al usar este kit, la empresa actúa mayoritariamente como "Desplegador" (Deployer). Si los flujos de trabajo se destinan a sectores críticos (recursos humanos, banca, educación), la empresa debe cumplir con obligaciones de transparencia y monitorización humana.
- Tras usarlo, he verificado que el kit permite el uso de modelos de "propósito general". Según la Ley de IA, si la empresa modifica sustancialmente el modelo para un uso de alto riesgo, podría ser considerada "Proveedor", asumiendo mayores responsabilidades técnicas y de certificación.
- Se debe garantizar la alfabetización en IA del personal que interactúe con los resultados generados por los modelos locales (Ollama).

Privacidad y protección de datos
- **Responsabilidades:** La empresa española es la Responsable del Tratamiento al 100%. n8n (la empresa) no tiene acceso a los datos procesados en una instancia self-hosted.
- **Ubicación de los datos:** Estrictamente donde se aloje el servidor (local o VPS). Para cumplimiento óptimo, el servidor debe estar en España o la UE.
- **Transferencia internacional:** No existen de forma nativa, lo cual es una ventaja competitiva legal. Cualquier nodo de n8n que conecte con servicios externos (ej. Slack, Google) reactivará la necesidad de evaluar estas transferencias.
- **Derechos ARCO:** La empresa debe implementar manualmente los mecanismos de acceso, rectificación y supresión dentro de las bases de datos incluidas (PostgreSQL/Qdrant).

Propiedad intelectual
- **Propiedad de datos:** La propiedad de los datos de entrenamiento (embeddings) y de entrada pertenece exclusivamente a la empresa.
- **Propiedad del resultado:** Según la legislación española, las obras generadas íntegramente por IA no tienen derechos de autor, pero la estructura de los "workflows" (flujos de n8n) son propiedad intelectual de la empresa como software o base de datos protegida.
- n8n se distribuye bajo una licencia "FairCode" (Fair Code License) que permite su uso gratuito pero limita su reventa como servicio gestionado.

Usos y prohibiciones
- **Usos admitidos:** Automatización de procesos internos, análisis de documentos privados, creación de agentes de soporte interno.
- **Usos prohibidos:** No se puede utilizar la versión Community de n8n para ofrecer un servicio comercial de automatización competidor (SaaS) sin licencia Enterprise. El uso para vigilancia biométrica o puntuación

social está prohibido por la AI Act.

Seguridad y certificaciones
- **Seguridad:** El kit no viene "securizado" por defecto para producción. La seguridad depende de la configuración de Docker y del cortafuegos del servidor.
- **Certificaciones:** n8n declara cumplimiento con SOC2 en sus servicios cloud, pero en la versión self-hosted la certificación debe ser obtenida y mantenida por la empresa española sobre su propia infraestructura.

Otros
Es vital diferenciar entre las licencias: mientras el Starter Kit (scripts de despliegue) es Apache 2.0, los binarios de n8n están sujetos a la licencia n8n FairCode, que prohíbe explícitamente el uso para fines comerciales de "hosting" de la propia herramienta.

Fuentes consultadas:
- https://github.com/n8n-io/self-hosted-ai-starter-kit/blob/main/LICENSE
- https://docs.n8n.io/hosting/starter-kits/ai-starter-kit/
- https://n8n.io/fair-code/
- https://www.aepd.es/es/guias-y-herramientas/guias/guia-adecuacion-al-rgpd-de-tratamientos-que-incorporan-ia